# Quantifying Registration Uncertainty with Sparse Bayesian Modelling

Loïc Le Folgoc, Hervé Delingette, Antonio Criminisi, Nicholas Ayache

# Quantifying Registration Uncertainty with Sparse Bayesian Modelling

Loïc Le Folgoc, Hervé Delingette, Antonio Criminisi, Nicholas Ayache

***Abstract*—We investigate uncertainty quantification under a sparse Bayesian model of medical image registration. Bayesian modelling has proven powerful to automate the tuning of registration hyperparameters, such as the trade-off between the data and regularization functionals. Sparsity-inducing priors have recently been used to render the parametrization itself adaptive and data-driven. The sparse prior on transformation parameters effectively favors the use of coarse basis functions to capture the global trends in the visible motion while finer, highly localized bases are introduced only in the presence of coherent image information and motion. In earlier work, *approximate* inference under the sparse Bayesian model was tackled in an efficient Variational Bayes (VB) framework. In this paper we are interested in the theoretical and empirical quality of uncertainty estimates derived under this approximate scheme vs. under the exact model. We implement an (asymptotically) exact inference scheme based on reversible jump Markov Chain Monte Carlo (MCMC) sampling to characterize the posterior distribution of the transformation and compare the predictions of the VB and MCMC based methods. The true posterior distribution under the sparse Bayesian model is found to be meaningful: orders of magnitude for the estimated uncertainty are quantitatively reasonable, the uncertainty is higher in textureless regions and lower in the direction of strong intensity gradients.**

***Index Terms*—Registration, Sparse Bayesian Learning, Uncertainty Quantification, MCMC, Reversible Jump, Automatic Relevance Determination.**

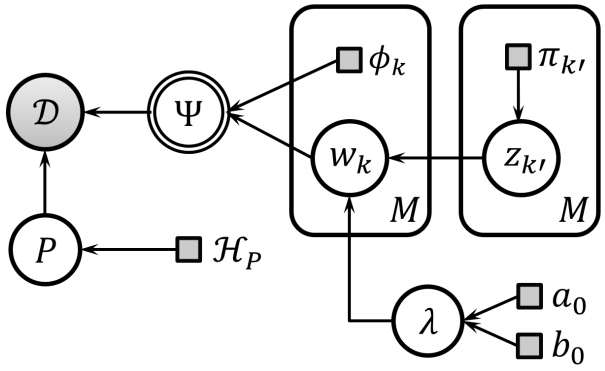


Fig. 1. Graphical model of registration. The generative model of data $D$ involves a transformation $\Psi$ of space, and noise governed by a set of underlying parameters $P$. Hyperpriors (with hyperparameters $\mathcal{H}_P$) are in turn imposed over the noise parameters. The transformation is parametrized as a linear combination of predefined basis functions $\{\phi_k,\, k = 1 \cdots M\}$ with associated weights $\boldsymbol{w}_k$. Priors on the transformation smoothness and on the relevance of individual bases introduce additional parameters ($\lambda$ and $z_{k'}$ respectively). Random variables are circled, hyperparameters are squared. Arrows capture conditional dependencies. Shaded nodes are observed variables or fixed hyperparameters. The transformation $\Psi$ is fully determined by its parent nodes (the $\phi_k$ and $\boldsymbol{w}_k$), hence the doubly circled node. The content of plates is replicated (M times).

## I. Introduction

Non-rigid image registration is an ill-posed task that supplements limited, noisy data with 'inexact but useful' prior knowledge to infer an optimal deformation between images of interest [1]. As a standard processing step in many pipelines for medical imaging, for computational anatomy & physiology, registration would benefit from the development of principled strategies to analyze its output and subsequently re-evaluate model assumptions. Bayesian modelling provides a framework to explicitly incorporate prior assumptions and reassess their relevance in retrospect. We focus here on another expected benefit of Bayesian approaches that is, the possibility to quantify uncertainty in the optimal solution.

Probabilistic approaches to registration and uncertainty quantification are not yet widespread in the literature. Gee and Bajcsy [2] laid the groundwork for a Bayesian interpretation of registration, extending the mechanical formulation of Broit [3]. Exploiting the Gaussian Markov random field structure inherited from a finite-element discretization of the domain, they characterize the posterior distribution of displacements by Gibbs sampling. Risholm *et al* [4] extend the approach to the case of unknown confidence on the observed data and on model priors respectively, aiming to address the critical issue of finding an objective trade-off between data fit and regularity-inducing priors. The so-called temperature hyperparameters are treated as latent variables and approximately marginalized over, while a Markov chain with full dimensional Metropolis-Hastings transitions traverses the space of transformation parameters. The aforementioned authors proceed in the framework of Markov Chain Monte Carlo (MCMC) sampling to explore the posterior distribution of model parameters. MCMC sampling yields an arbitrarily good characterization of the posterior provided that enough samples can be drawn within the available computational budget – inference becomes exact in non finite time. In practice, the computational burden and the technicality of the Markov chain implementation quickly become limiting factors. As an alternative, Variational Bayes (VB) inference provides tools to efficiently approximate the (true) posterior distribution on a chosen family of variational (approximate) posteriors. The choice of variational posterior realizes a trade-off between the computational burden and the quality of the estimates. Using

L. Le Folgoc is with the Asclepios Research Project, Inria Sophia Antipolis and with the Microsoft–Inria Joint Centre, France. H. Delingette and N. Ayache are with the Asclepios Research Project, Inria Sophia Antipolis, 2004 route des Lucioles BP 93, 06902, Sophia Antipolis, France.

A. Criminisi is with the Machine Learning and Perception Group, Microsoft Research Cambridge, United Kingdom.

a parametric (FFD) representation of the displacement field, Simpson *et al* [5], [6] approximate the posterior distribution within a 'convenient' family for which transformation parameters and model hyperparameters factorize. The variational factorization renders the approach applicable to real scale registration tasks. As a drawback estimates of uncertainty quantify variability in the displacement field *conditionally* to the inferred hyperparameters, but disregard uncertainty induced by hyperparameter variability. Although uncertainty quantification is peripheral to their work, Richard *et al* [7] develop for the related task of atlas building a mixed SAEM and MCMC approach where nodes of the finite-element mesh are updated via Metropolis-Hastings-Within-Gibbs transitions; Zhang *et al* [8] implement a mixed SAEM and Hybrid Monte Carlo approach for a Bayesian MAP estimation of the template and of temperature hyperparameters in a diffeomorphic setting.

In this paper, we compare the approximate posterior returned by a Variational Bayes method with an MCMC method based on the same underlying model. We focus on the Bayesian model of registration developed in earlier work [9]. The main goal of this model is to allow not only for the automatic determination of registration parameters (such as the trade-off between image similarity and regularization functionals), but also for a data-driven, multiscale, spatially adaptive parametrization of deformations via the recourse to a sparsity-inducing prior on transformation parameters.

Our contribution is twofold. Firstly, the complexity of the model renders inference non trivial. While in our previous work approximate inference was conducted on the grounds of Variational Bayes, we adopt here an exact MCMC-based approach. At a high level, the space of transformation parameters is explored by a reversible jump Markov chain [10]. It provides a principled mechanism to elegantly jump between competing parametrizations of the displacement field, regardless of their dimensionalities, without the prohibitely expensive computation of so-called Bayes factors. This allows to seamlessly refine the parametrization of the transformation, adapting the granularity of the parametrization to the granularity of the underlying motion and the local informativeness of the image, all the while exploring the most likely deformations. At a lower level, we capitalize on closed form marginalization of most nuisance variables, and integrate second-order knowledge of the posterior distribution in proposal kernels. This yields an algorithm that reliably and consistently traverses the parameter space towards the most likely deformations in spite of the model intricacies.

Secondly, we compare the expectation and uncertainty predicted by both the fast (approximate) Variational Bayes inference and the (asymptotically) exact MCMC inference scheme both on empirical and theoretical grounds. We found that the expectation is typically well approximated by the VB inference, but that the uncertainty is underestimated. We exhibit two mechanisms that explain this behaviour. Furthermore we show that uncertainties predicted by the exact model are consistent with intuition: the orders of magnitude are sound, the uncertainty is higher in textureless regions and lower in the direction of strong intensity gradients.

The article unfolds as follows. In part II we describe the sparse Bayesian model of registration and devise a principled strategy for exact inference. The proposed design of the Markov chain exploits insight gained about the model to bypass standard impediments of MCMC schemes. Hyperparameter uncertainty is fully accounted for by marginalization of the nuisance variables. In part III we review breakdown scenarii in which the approximate posterior significantly departs from the true posterior, leading to poor approximate predictive uncertainty. In part IV we conduct preliminary experiments to assess the validity of MCMC uncertainty estimates.

## II. Statistical Model and Inference

Registration infers, from prior knowledge and limited data $\mathcal{D}$, a transformation of space $\Psi$ that pairs homologous features in objects of interests (*e.g.* organs or vessels, in a medical setting). The section starts with a succinct description of the registration model, and offers insight into its mechanisms. Fig. 1 provides a graphical representation thereof. An MCMC approach for systematic characterization of the posterior distribution is then devised.

### A. Bayesian Model of Registration

*1) Likelihood model:* The generative model of data makes explicit the relationship between the data $\mathcal{D}$ and the spatial mapping $\Psi$. It is specified by a likelihood model $p(\mathcal{D}|\Psi;P)$ (often conditioned on a set of hyperparameters $P$) that typically assumes the form of a Boltzmann distribution $p(\mathcal{D}|\Psi;P) \propto \exp -\mathcal{E}_{\mathcal{D}}(\mathcal{D},\Psi;P)$. For landmark registration, a transformation that approximately maps corresponding key points $\{t_i\}$ and $\{T_i\}$, $i = 1 \cdots N$, between a template object and a target object is sought. A standard choice of energy is the sum of squared distances between pairings, up to multiplicative factor:

$$\mathcal{E}_{\mathcal{D}}(\mathcal{D},\Psi;\beta) = \frac{\beta}{2}\sum_{i=1}^{N} \|T_i - \Psi(t_i)\|^2 . \tag{1}$$

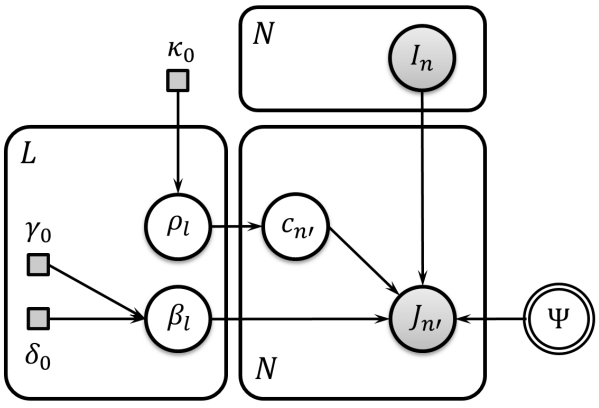


Fig. 2. Graphical representation of the generative data model (using the same graphical symbols as in Fig. 1). Residuals between the fixed image $J$ and the warped image $I \circ \Psi^{-1}$ are assumed to be distributed according to a mixture of $L$ Gaussian components whose parameters $\rho_l$ (probability of falling in the $l$th component) and $\beta_l$ (inverse variance a.k.a. precision parameter for the $l$th Gaussian component) are regarded as latent variables. $c_{n'} \in \{1 \cdots L\}$ assigns the corresponding voxel to one of the $L$ mixture components.

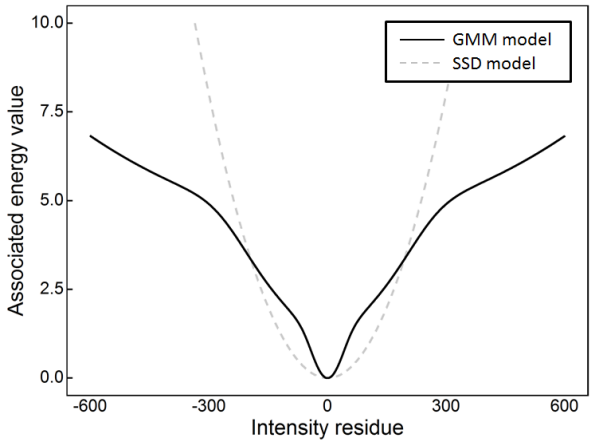


Fig. 3. Energies corresponding to example GMM (black) and SSD (dotted) models. The voxelwise penalty is shown as a function of the intensity residual. The effective soft threshold on the penalty incurred for large intensity residuals results in increased robustness of the GMM.

For pairwise registration of a fixed image $J$ and a moving image $I$, a mixture-of-Gaussians model (GMM) of intensity residuals is adopted here as a flexible and robust variant of the widespread sum of squared differences (SSD). Fig. 2 summarizes this model of data in graphical form, making explicit the nodes $P, \mathcal{H}_P, \mathcal{D}$ of Fig. 1. The multiple components of the GMM naturally cope with the fact that intensity residuals may rightfully take high outlier values for an undeterminate fraction of voxels, because of acquisition artefacts, heteroscedastic noise and model inaccuracies. At voxel center $v_i$, the intensity residual $r_i = J(v_i) - I[\Psi^{-1}(v_i)]$ is assigned to the $l$th component of the mixture, $1 \leq l \leq L$, if the $L$-way categorical variable $c_i \in \{1 \cdots L\}$ takes value $l$. If so the residual $r_i$ follows a normal distribution $\mathcal{N}(0, \beta_l^{-1})$. The component assignment $c_i$ follows a categorical distribution and takes value $l$ with probability $\rho_l$, normalized such that $\sum_{l=1}^{L} \rho_l = 1$. For distinct voxels $v_i$ and $v_j$, residuals $r_i$ and $r_j$ (resp. component assignments $c_i$ and $c_j$) are assumed to be independent. The corresponding GMM energy $\mathcal{E}_{\mathcal{D}}(\mathcal{D}, \Psi; \boldsymbol{\beta}, \boldsymbol{\rho})$ is given by Eq. (2), with $Z_l = \sqrt{2\pi/\beta_l}$ a normalizing constant:

$$-\sum_{i=1}^{N} \log \sum_{l=1}^{L} \frac{\rho_l}{Z_l} \exp -\frac{\beta_l}{2} \left( J[v_i] - I[\Psi^{-1}(v_i)] \right)^2 \quad (2)$$

Fig. 3 shows the typical profile of the GMM energy comparatively with the SSD. The assumption of independence of voxelwise residuals is known not to hold (see *e.g.* [5], [9]) and to affect the outcome of the probabilistic registration. Since a proper probabilistic account of correlations in intensity residuals is both beyond the scope of this work and irrelevant to the ensuing developments, the *Virtual Decimation* scheme of [5] is reproduced instead for simplicity.

*2) Transformation parametrization:* A small deformation standpoint is adopted for convenience. The displacement field $\mathrm{u}\colon \mathrm{x} \in \Omega \subset \mathbb{R}^d \mapsto u(x) = \Psi^{-1}(x) - x \in \mathbb{R}^d$ is parametrized by a linear combination of $M$ basis functions $\phi_k(.)$ with associated weight $w_k \in \mathbb{R}^d$:

$$u(x) = \sum_{1 \leq k \leq M} \phi_k(x) w_k = \boldsymbol{\phi}(x)^\intercal \boldsymbol{w} \,. \quad (3)$$

$\boldsymbol{\phi}(\mathrm{x}) = \left(\phi_1(\mathrm{x}) \cdots \phi_M(\mathrm{x})\right)^\intercal$ and $\boldsymbol{w}^\intercal = \left(w_1^\intercal \cdots w_M^\intercal\right)$ are respectively the concatenation, for $k = 1 \cdots M$, of $\phi_k(\mathrm{x})$ and $w_k$. Arbitrary choices of basis functions $\phi_k$ are possible. B-splines (*e.g.* [11]) present desirable properties in terms of smoothness and interpolation. Here the $\phi_k$'s instead consist of multiscale Gaussian radial basis functions (RBFs) whose centers lie on a regular grid of points (typically, decimated voxel centers). Multiscale Gaussian RBFs possess attractive analytical and computational properties.

*3) Transformation priors:* The weights $\boldsymbol{w}$ are endowed with a generalized Spike-&-Slab prior that favours both smoothness of the resulting displacement field and sparsity in its parametrization. The properties of this prior are central to the proposed 'sparse Bayesian' modelling and to our analysis thereof. Each basis $\phi_k$ is assigned a distinct activation variable $z_k$ that controls its inclusion in the active parametrization (or exclusion therefrom). If $z_k = 0$ the basis $\phi_k$ is pruned out of the active parametrization. We do so by designing $p(w_k | z_k = 0)$ as a Dirac distribution centered at 0. If $z_k = 1$ the basis $\phi_k$ is included in the parametrization. The prior on such bases is designed as a joint, structured Gaussian distribution that penalizes lack of smoothness in the induced displacement field [12]. Let us denote by $\mathcal{S}$ the set of such indices $k$ for which $z_k = 1$ and by $\boldsymbol{w}_{\mathcal{S}}$ the concatenation of the corresponding subset of weights $\{w_k,\, k \in \mathcal{S}\}$. For an arbitrary linear differential operator $D$, we wish to penalize high values of the quadratic energy $\|D\mathrm{u}\|^2 = \boldsymbol{w}_{\mathcal{S}}^\intercal \mathbf{R}_{\mathcal{S}} \boldsymbol{w}_{\mathcal{S}}$, where $\mathbf{R}_{\mathcal{S}}$ is the $|\mathcal{S}| \times |\mathcal{S}|$ matrix whose $k,l$-th coefficient is $\langle D\phi_k | D\phi_l \rangle$. The Gaussian distribution $\mathcal{N}(\boldsymbol{w}_{\mathcal{S}} \,|\, \mathbf{0}, \{\lambda\, d|\mathcal{S}|\}^{-1} \mathbf{R}_{\mathcal{S}}^{-1})$ is a natural choice of prior for $p(\boldsymbol{w}_{\mathcal{S}}|\mathcal{S})$, that we adopt henceforth. Note that the covariance normalization by $d|\mathcal{S}|$, where $d$ is the image dimension, departs from that of [9]. Under this prior $\lambda\, d|\mathcal{S}| \cdot \boldsymbol{w}_{\mathcal{S}}^\intercal \mathbf{R}_{\mathcal{S}} \boldsymbol{w}_{\mathcal{S}}$ is $\chi^2(d|\mathcal{S}|)$ distributed so that $\lambda$ immediately relates to the expectation of the energy: $\mathbb{E}_{p(\boldsymbol{w}_{\mathcal{S}}|\mathcal{S})}(\|D\mathrm{u}\|^2) = \lambda^{-1}$ and $\mathbb{E}_{p(\boldsymbol{w})}(\|D\mathrm{u}\|^2) = \lambda^{-1}$. The prior over all weights $\boldsymbol{w}$ conditioned on the state of the gate variables $\boldsymbol{z} = \left(z_1 \cdots z_M\right)^\intercal$ is best summarized in the form of Eq. (4), where $-\mathcal{S}$ is the complement of $\mathcal{S}$:

$$p(\boldsymbol{w}|\boldsymbol{z}, \lambda) = \mathcal{N}(\boldsymbol{w}_{\mathcal{S}} \,|\, \mathbf{0}, \frac{1}{\lambda\, d|\mathcal{S}|} \mathbf{R}_{\mathcal{S}}^{-1}) \cdot \mathcal{N}(\boldsymbol{w}_{-\mathcal{S}} \,|\, \mathbf{0}, \mathbf{0}) \,. \quad (4)$$

*4) Hyperpriors:* Parameters introduced in the specification of priors are in turn treated as latent variables. $\lambda$ is endowed with a Gamma prior $\Gamma(\lambda | a_0, b_0)$ that is conjugate to $p(\boldsymbol{w}|\boldsymbol{z}, \lambda)$. The parameters $\beta_l$ (resp. $\beta$) involved in the likelihood model for image (resp. landmark) registration are endowed with independent Gamma priors $\Gamma(\beta_l | \gamma_0, \delta_0)$. The noise mixture proportions $\boldsymbol{\rho} = \{\rho_1 \cdots \rho_L\}$ are assigned a Dirichlet prior $\mathrm{Dir}(\boldsymbol{\rho}|\boldsymbol{\kappa})$, with $\boldsymbol{\kappa} = (\kappa_1 \cdots \kappa_L)$.

Independent Bernoulli priors $\mathcal{B}(z_k|\pi_k)$ on each $z_k$ constitute a natural, conjugate hyperprior specification for the activation variables $\boldsymbol{z}$. The positive mass $1 - \pi_k$ concentrated at $w_k = 0$ as a result explicitly encodes sparsity. Assuming all $\pi_k = \pi_0$ to be equal, all parametrizations using the same number of

active bases $|\mathcal{S}|$ are a priori equally probable. In addition the cost of including a new basis in the active parametrization is independent of the current number of active bases. However, we opt instead for a stronger prior, $p(\boldsymbol{z}) \propto \Gamma(\frac{d|\mathcal{S}|}{2})^{-1}$. The Gamma function $\Gamma(\cdot)$ is a natural extension of the (integer) factorial to real values, yielding a prior that increasingly penalizes each new inclusion. This prior was found to perform better w.r.t. sparsity, as can be theoretically argued from the analysis of the marginal prior $p(\boldsymbol{w}|\boldsymbol{z})$.

### B. Model analysis

*1) Marginal prior and marginal likelihood:* Critical insight into the statistical model can be gained by considering the prior $p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H})$ and likelihood $p(\mathcal{D}|\boldsymbol{w},\boldsymbol{c},\mathcal{H})$ with so called *temperature* parameters $\lambda$ and $\boldsymbol{\beta}$ marginalized over, *e.g.*:

$$p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H}) = \int_{\mathbb{R}_+} p(\boldsymbol{w}|\boldsymbol{z},\lambda,\mathcal{H})p(\lambda|\mathcal{H})d\lambda\,. \tag{5}$$

The multivariate Student distribution $t_\nu(\cdot|\boldsymbol{\mu},\boldsymbol{\Lambda})$ with location parameter $\boldsymbol{\mu}$, inverse scale matrix $\boldsymbol{\Lambda}$ and $\nu$ degrees of freedom naturally appears in analytic derivations, yielding the following expressions for the prior and likelihood:

$$p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H}) = \mathcal{N}\left(\boldsymbol{w}_{-\mathcal{S}}\middle|\boldsymbol{0},\boldsymbol{0}\right) t_{\nu_\lambda}\left(\boldsymbol{w}_{\mathcal{S}}\middle|\,\boldsymbol{0},\frac{a_0}{b_0}\,d|\mathcal{S}|\,\mathbf{R}_{\mathcal{S}}\right) \tag{6}$$

$$p(\mathcal{D}|\boldsymbol{w},\boldsymbol{c},\mathcal{H}) = \prod_{l=1}^{L} t_{\nu_l}\left(\boldsymbol{I}_l\circ\Psi_{\boldsymbol{w}}^{-1}\middle|\,\boldsymbol{J}_l,\frac{\gamma_0}{\delta_0}\mathbf{I}\right) \tag{7}$$

where $\nu_\lambda = 2a_0$, $\nu_l = 2\gamma_0$, $\mathcal{S}$ is the set of active bases and $|\mathcal{S}| = \sum_k z_k$ its cardinal. $\boldsymbol{J}_l = \left(\cdots J[v_i]\cdots\right)^\intercal_{i|c_i=l}$ is the vector of voxel values in image $J$, for those voxels assigned to component $l$, and $\boldsymbol{I}_l\circ\Psi_{\boldsymbol{w}}^{-1} = \left(\cdots I[\Psi_{\boldsymbol{w}}^{-1}(v_i)]\cdots\right)^\intercal_{i|c_i=l}$ is similarly defined for the warped image $I\circ\Psi_{\boldsymbol{w}}^{-1}$. For a fixed choice of active bases $\boldsymbol{z}$, the posterior distribution of the weights $p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{c},\mathcal{D},\mathcal{H})$ is proportional to the product of the prior Eq. (6) and likelihood Eq. (7). In the limit of uninformative hyperpriors $a_0,\gamma_0\to 0$, $\beta_0,\delta_0\to 0$ and assuming $L=1$ for the sake of illustration,

$$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{c},\mathcal{D},\mathcal{H}) \propto \mathcal{N}\left(\boldsymbol{w}_{-\mathcal{S}}\middle|\boldsymbol{0},\boldsymbol{0}\right)\,\frac{1}{\chi_{\text{lik}}[\boldsymbol{w}]^N}\frac{1}{\chi_{\text{pr}}[\boldsymbol{w}]^{d|\mathcal{S}|}}\,. \tag{8}$$

where $\chi_{\text{lik}}[\boldsymbol{w}]^2$ is the data error and $\chi_{\text{pr}}[\boldsymbol{w}]^2 = \|D\mathrm{u}_{\boldsymbol{w}}\|^2$ the regularizing energy. In particular the posterior distribution is invariant to rescaling of the data error, and hence to rescaling of the intensity profile, after marginalizing over temperature parameters. Note also that, for a fixed parametrization $\boldsymbol{z}$, the ratio of posterior probabilities of two distinct parameter sets $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ may become arbitrarily overwhelmed by the prior as the number of bases in the parametrization grows ($|\mathcal{S}|\gg N$). If not for sparsity, this might render MCMC characterization of the posterior unreliable (using *e.g.* Metropolis Hastings transitions), potentially making its outcome dependent on the size of the parametrization. Fortunately the proposed sparse model has a clear mechanism to prevent overparametrization and render overlapping bases largely mutually exclusive, as discussed next.

*2) Prior probability of basis inclusion:* Interactions between overlapping bases can be better understood by looking at the probability $p(z_k|\boldsymbol{w}_{-k},\boldsymbol{z}_{-k},\mathcal{H})$ of inclusion of a new basis $z_k$ given a known configuration $\boldsymbol{z}_{-k}$ for the other bases and their associated weights $\boldsymbol{w}_{-k}$. The state $\boldsymbol{w}_{-k}$ of other bases informs us about the expected regularity of the signal $\mathrm{u}_{\boldsymbol{w}}$, introducing dependencies between $z_k$ and $\boldsymbol{z}_{-k}$ conditionally to $\boldsymbol{w}_{-k}$. Denoting by $\tilde{\boldsymbol{z}}$ (resp. $\boldsymbol{z}$) the state with $z_k=1$ (resp. $z_k=0$), we see from Bayes' rule that:

$$\frac{p(z_k=1|\boldsymbol{w}_{-k},\boldsymbol{z}_{-k})}{p(z_k=0|\boldsymbol{w}_{-k},\boldsymbol{z}_{-k})} = \frac{p(\boldsymbol{w}_{-k}|\tilde{\boldsymbol{z}})}{p(\boldsymbol{w}_{-k}|\boldsymbol{z})}\frac{p(\tilde{\boldsymbol{z}})}{p(\boldsymbol{z})} \tag{9}$$

where the dependence on hyperparameters is made implicit for convenience of notations. Leaving details of derivations aside, we note that in the limit of uninformative values, the ratio of Eq. (9) takes the form of

$$\frac{p(\tilde{\boldsymbol{z}})}{p(\boldsymbol{z})}\,\left(\frac{|\kappa_k|}{|R_{k,k}|}\right)^{1/2}\left(1-\frac{\mu_{\text{pr}}^{k\intercal}\mathrm{R}_{k,k}\mu_{\text{pr}}^k}{\boldsymbol{w}_{-k}^\intercal\mathbf{R}_{\mathcal{S}}\boldsymbol{w}_{-k}}\right)^{-\frac{d|\mathcal{S}|}{2}} \tag{10}$$

where $\mathcal{S}$ is the set of active bases (excluding $k$), $\mu_{\text{pr}}^k = -\mathrm{R}_{k,k}^{-1}\mathbf{R}_k^\intercal\boldsymbol{w}_{-k}$ and $\kappa_k = \mathrm{R}_{k,k} - \mathbf{R}_k^\intercal\mathbf{R}_{\mathcal{S}}^{-1}\mathbf{R}_k$. The middle factor penalizes the inclusion of basis $k$ if it overlaps with bases in the active set $\mathcal{S}$, in the sense of the metric induced by $\mathbf{R}$. $\kappa_k$ is a measure of overlap of basis $k$ with all bases in the active set $\mathcal{S}$ and is null if basis $k$ is perfectly collinear to $\mathcal{S}$. The rightmost factor favors the inclusion of basis $k$ if it is a priori expected to yield a significant increase in regularity.

### C. Posterior Exploration by MCMC Sampling

For any set of points $X = \{x_1\cdots x_n\}$ in the admissible domain $\Omega$, consider the vector of displacements $\boldsymbol{u}_X^\intercal = \left(u(x_1)^\intercal\cdots u(x_n)^\intercal\right)$. We wish to characterize the joint posterior distribution $p(\boldsymbol{u}_X|\mathcal{D},\mathcal{H})$ of any such vector of displacements for any discrete set $X$. To that aim we merely need to characterize the posterior distribution $p(\boldsymbol{w}|\mathcal{D},\mathcal{H})$ of the weights $\boldsymbol{w}$ involved in the parametrization of the transformation $\Psi^{-1}$ sufficiently well.

*1) Related work:* MCMC methods are tools of predilection to explore arbitrarily complex distributions in a principled manner. Gibbs sampling [13] cycles between latent variables, sampling from their conditional distributions in turn while other model variables remain fixed. It is attractive when conditional distributions are known in closed form whereas the joint distribution is untractable or computationally costly to sample. When the conditional cannot be sampled directly, a component-wise proposal may be used instead within a Metropolis-Hastings (MH) step (Metropolis-Within-Gibbs). Unfortunately, Gibbs sampling of temperature parameters is prone to failure, with the chain drifting away from regions of high probability for the duration of any finite MCMC run. Collapsing temperature parameters $\lambda$, $\boldsymbol{\beta}$ when sampling regressor variables $\boldsymbol{w}$ is highly opportune. In the context of registration, Risholm et al. [4] propose a MH scheme where marginalizing over temperature parameters induces the expensive computation of partition functions, for which an intricate procedure based on Laplace approximations is designed. In the proposed model, the computation of partition

functions (specifically, marginal likelihoods, a.k.a. *evidences*) may arise as well when sampling gate variables $z_k$. Selecting a specific configuration $\boldsymbol{z}$ can be interpreted as a choice between competing models of varying complexity and dimensionality. The problem of estimating the evidence for a model is well studied in the statistical literature. A variety of methods exist, ranging from the straightforward Laplace approximation to more principled approaches typically exploiting samples from the (possibly augmented) posterior, including Chib's method [14], importance sampling, bridge sampling, path sampling (see *e.g.* [15]) and reversible jump MCMC [10]. The latter approach is in fact primarily concerned with sampling from a posterior distribution involving competing models (freely jumping between models in the process) and merely obtains evidence ratios as a byproduct. Reversible jump MCMC is appealing in our setting where competing models $\boldsymbol{z}$ are organized in series of nested models of increasing complexity, rendering its machinery mostly invisible. Reversible jump MCMC proceeds in the general framework of Metropolis-Hastings, hence a sound proposal must be crafted. We derive a sensible family of proposals from a modal analysis of the posterior distribution.

*2) Modal analysis of the posterior & proposal:* For the model described in II-A, the Laplace approximation of the (conditional) posterior $p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{z}, \boldsymbol{c}, \mathcal{H})$ around its mode $\boldsymbol{w}_* = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{z}, \boldsymbol{c}, \mathcal{H})$ takes the following form:

$$
\begin{aligned}
-\log p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{z}, \boldsymbol{c}, \mathcal{H}) \approx \\
\frac{1}{2}\frac{\gamma_0 + N/2}{\delta_0 + \chi^2_{\text{lik}}/2}\,(\boldsymbol{T}_* - \boldsymbol{\Phi}\boldsymbol{w})^\intercal \mathbf{H}_*(\boldsymbol{T}_* - \boldsymbol{\Phi}\boldsymbol{w}) + \\
\frac{1}{2}\frac{a_0 + |\mathcal{S}|/2}{b_0 + \chi^2_{\text{pr}}|\mathcal{S}|/2}\,d|\mathcal{S}|\,\boldsymbol{w}^\intercal \mathbf{R}_{\mathcal{S}}\boldsymbol{w} + \text{const}\,. \qquad (11)
\end{aligned}
$$

where for the sake of illustration we take a single component mixture ($L = 1$, $c_i = 1$ for all $i$, $\boldsymbol{\beta} = \beta$). $\chi^2_{\text{pr}} = \boldsymbol{w}_*^\intercal \mathbf{R}_{\mathcal{S}} \boldsymbol{w}_*$ is the energy in the displacement field, $\chi^2_{\text{lik}}$ is the data error $\chi^2_{\text{lik}} = \sum_{i=1}^{N} (J[v_i] - I[\Psi_*^{-1}(v_i)])^2$ and we discard higher order terms in $b_0$, $\delta_0$. $\boldsymbol{T}_*^\intercal = \begin{pmatrix} T_{1*}^\intercal & \cdots & T_{N*}^\intercal \end{pmatrix}$ is a set of *virtual* pairings whose value does not depend on $\boldsymbol{\beta}$, $\lambda$. $\mathbf{H}_*$ is a block diagonal matrix whose $i$th diagonal block $\mathbf{H}_i^*$ is the $d \times d$ precision matrix associated to the $i$th virtual pairing $T_{i*}$. The factors stemming from the marginalization:

$$
\beta_* = \frac{\gamma_0 + N/2}{\delta_0 + \chi^2_{\text{lik}}/2}\,, \quad \lambda_* = \frac{a_0 + |\mathcal{S}|/2}{b_0 + \chi^2_{\text{pr}}|\mathcal{S}|/2} \qquad (12)
$$

are commensurable to temperature parameters. The approximation of the conditional posterior is Gaussian (Eq. (11) is quadratic) and admits the more obvious canonical form $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^\intercal(\beta_* \mathbf{H}_*)\boldsymbol{T}_*$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^\intercal \beta_* \mathbf{H}_* \boldsymbol{\Phi} + \lambda_* |\mathcal{S}| \mathbf{R}_{\mathcal{S}})^{-1}$. The Laplace approximation provides a reasonable approximation of the posterior and a judicious starting point to design proposals. Component-wise proposals that leave most of the activation variables $z_l$ and the corresponding weights $w_l$ unchanged will be of particular interest to us (cf. section II-C3). A natural idea is to use the conditionals $\tilde{w}_k \sim \mathcal{N}(\mu^k_{\text{pos}}, \Sigma_k)$ of the Laplace approximation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as proposal distributions. Because they neither require the actual computation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ nor involve inner products $\phi_k^\intercal(\beta_* \mathbf{H}_*)\phi_l$, these 'Gibbs-like' proposals are computationally appealing. As a final tweak to alleviate modal assumptions, we reintroduce dependency on the current value of $w_k$, yielding the following component-wise proposal instead, with $0 \leq r_{\text{HMALA}} \leq 1$ and $s \geq 1$:

$$
q_k(w_k \to \tilde{w}_k) = \mathcal{N}\left(\tilde{w}_k \,|\, m_k(w_k), s\Sigma_k\right) \qquad (13)
$$

$$
m_k(w_k) = (1 - r_{\text{HMALA}})\, w_k + r_{\text{HMALA}}\, \mu^k_{\text{pos}} \qquad (14)
$$

If not set to 1, the factor $s$ accounts for potentially fatter tails of the true conditional posterior in the proposal. $\mu^k_{\text{pos}}$ and $\Sigma_k$ depend on $\mathbf{H}_*$ and $\boldsymbol{T}_*$, which in the formal reasoning based on the Laplace approximation are computed around $\Psi_*^{-1}(\cdot) = \text{Id} + \boldsymbol{\phi}(\cdot)^\intercal \boldsymbol{w}_*$. In fact $\boldsymbol{T}_*$ and $\mathbf{H}_*$ can be replaced by $\boldsymbol{T}_{\boldsymbol{w}}$ and $\mathbf{H}_{\boldsymbol{w}}$ computed from a (local) quadratic approximation of $p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{z}, \boldsymbol{c}, \lambda_*, \beta_*)$ around the current $\Psi^{-1}(\cdot) = \text{Id} + \boldsymbol{\phi}(\cdot)^\intercal \boldsymbol{w}$. In that case Eq. (13), (14) exactly coincide with a component-wise Hessian preconditioned Metropolis Adjusted Langevin Algorithm (HMALA) [16]–[18], which exploits first and second order local information about the target distribution for increased efficiency. However the local approximation generates additional computations at each step and offers little gain if we expect the posterior to be unimodal. Given our experimental settings, we use the global approximation with adaptation during the burn-in phase (at that stage $\lambda_*$, $\beta_*$, $\boldsymbol{T}_*$ and $\mathbf{H}_*$ are recomputed every few iterations from statistics $|\mathcal{S}|$, $\chi^2_{\text{pr}}$, $\chi^2_{\text{lik}}$ averaged with decaying weights over past samples).

*3) Reversible jump MCMC scheme:* The groundwork for this scheme was laid in sections II-B1, II-B2, II-C2. The reversible jump procedure itself lets us generate samples of the joint posterior $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})$ with temperature parameters marginalized over. Dropping irrelevant variables in the generated samples, we obtain samples of the marginals of interest, *e.g.* $p(\boldsymbol{w}|\mathcal{D}, \mathcal{H})$. The reversible jump scheme simply proposes to move from a current state $\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}$ to a new state $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{c}}$ and computes a Metropolis-Hastings acceptance ratio for the proposal, leading to acceptance or rejection of the new state. For the sake of simplicity, proposals for a new state of $\boldsymbol{w}, \boldsymbol{z}$ may be made separately from those of $\boldsymbol{c}$. For the latter, the most natural proposal exactly results in collapsed Gibbs sampling of each $c_i$, see *e.g.* [19][1]. For $\boldsymbol{w}, \boldsymbol{z}$ we design basic moves that – when combined – allow to add, remove or switch active bases as well as update several components of $\boldsymbol{w}$. These basic moves are combined to craft proposal distributions $Q(\boldsymbol{w}, \boldsymbol{z} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}})$ for which the probability of a move $\boldsymbol{w}, \boldsymbol{z} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$ has direct symmetries with that of the reverse move $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}} \to \boldsymbol{w}, \boldsymbol{z}$, so that the acceptance ratio

$$
\min\left(1, \frac{p(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})}{p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})} \frac{Q(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c} \to \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c})}{Q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c})}\right) \qquad (15)
$$

becomes particularly straightforward to compute. The basic moves are:

*a) Basis removal*. For a basis $k$ such that $z_k = 1$, set $\tilde{z}_k = 0$ and $\tilde{w}_k = 0$. The symmetric move is the basis addition.

*b) Component-wise update*. For a basis $k$ such that $z_k = 1$, propose a new $\tilde{w}_k \sim q_k(w_k \to \tilde{w}_k)$ according to Eq. (13),

[1] A complete and concise summary of the relevant derivations and schemes is given in http://www.kamperh.com/notes/kamper_bayesgmm13.pdf

**Algorithm 1:** Proposal $Q_k(\boldsymbol{w}, \boldsymbol{z}, \text{reverse\_traversal} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \text{reverse\_traversal}^*)$.

$n_{\text{neighb}}$ is an integer fixed in advance.
Set $\tilde{\boldsymbol{w}} = \boldsymbol{w}$, $\tilde{\boldsymbol{z}} = \boldsymbol{z}$.

Draw one of 3 competing events: ***on-off***, ***exchange***, ***update***.

**if** $\phi_k$ **inactive** *and* **update then**
  **Exit**. No action to implement (as $\tilde{w}_k = 0$).

**if exchange** *and* $\phi_k$ **active then**
  Draw an inactive basis $\phi_{k^*}$ to replace $\phi_k$. A proposal that favours well-aligned bases is designed.
**else if exchange** *and* $\phi_k$ **inactive then**
  Draw an active basis $\phi_{k^*}$ to replace $\phi_k$. A proposal that favours well-aligned bases is designed.

**if** $\phi_k$ **active** *and* **on-off** *or* **exchange then**
  Set $\tilde{z}_k = 0$ and $\tilde{w}_k = 0$.

**if on-off** *or* **update then**
  - For ***update***: set $\mathcal{I} = \{k\}$.
  - For ***on-off***: Set $\mathcal{I} \subset \mathcal{S} \backslash \{k\}$ to a list of $n_{\text{neighb}}$ active bases, favouring bases well-aligned with $\phi_k$.
  - If reverse_traversal $= 1$, reverse the ordering of $\mathcal{I}$.

  **for** $l \in \mathcal{I}$ **do**
    $\tilde{w}_l^{\text{new}} \sim q_l(\tilde{\boldsymbol{w}} \to \tilde{\boldsymbol{w}}^{\text{new}})$ and $\tilde{w}_l = \tilde{w}_l^{\text{new}}$

**if** $\phi_k$ **inactive** *and* **on-off then**
  Set $\tilde{z}_k = 1$ and $\tilde{w}_k \sim q_k(w_k \to \cdot)$, using $r_{\text{HMALA}} = 1$.
**else if** $\phi_k$ **inactive** *and* **exchange then**
  Set $\tilde{z}_{k^*} = 1$ and $\tilde{w}_{k^*} \sim q_{k^*}(w_{k^*} \to \tilde{w}_{k^*})$ ($r_{\text{HMALA}} = 1$).

**if on-off** *or* **exchange then**
  Switch the state of the binary variable reverse_traversal.

(14) with a fixed $0 \leq r_{\text{HMALA}} \leq 1$. This move is its own symmetric (using the reverse update).

*c) Basis addition.* For a basis $k$ such that $z_k = 0$, set $\tilde{z}_k = 1$ and propose a new $\tilde{w}_k$ according to Eq. (13), (14) with $r_{\text{HMALA}} = 1$. The symmetric move is the basis removal.

The family of proposals $Q_k(\cdot)$ that we design combines these basic moves in such a way that when reversed, the sequence of moves induced by the proposal $Q_k(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c})$ coincides exactly with the sequence of moves induced by $Q_k(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c} \to \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c})$. The proposal and reverse proposal travel along the same path in opposite directions, drastically reducing the computational load when evaluating Eq. (15). Each proposal $Q_k$ revolves primarily around the corresponding basis $\phi_k$ and is defined as per Algorithm 1 (where we introduced a binary variable reverse_traversal to address technicalities). Using $Q_k$, we define a transition kernel $P_k$ conventionally: given the current state $\boldsymbol{w}_t, \boldsymbol{z}_t, \boldsymbol{c}_t$, we propose a new state $\tilde{\boldsymbol{c}} = \boldsymbol{c}_t$, $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}} \sim Q_k(\boldsymbol{w}_t, \boldsymbol{z}_t \to \cdot)$. The state is accepted with probability given by Eq. (15), in which case we set $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}, \boldsymbol{c}_{t+1}) = (\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{c}})$; otherwise we stay at the current state and $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}, \boldsymbol{c}_{t+1}) = (\boldsymbol{w}_t, \boldsymbol{z}_t, \boldsymbol{c}_t)$. Computation of the acceptance ratio is relatively straightforward by construction, since the ratio of posterior probabilities involved in Eq. (15) can be rewritten as:

$$\frac{p(\mathcal{D}|\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c}, \mathcal{H})}{p(\mathcal{D}|\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}, \mathcal{H})} \cdot \frac{p(\tilde{\boldsymbol{w}}|\tilde{\boldsymbol{z}}, \boldsymbol{c}, \mathcal{H}) p(\tilde{\boldsymbol{z}}|\mathcal{H})}{p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{c}, \mathcal{H}) p(\boldsymbol{z}|\mathcal{H})} \tag{16}$$

The leftmost factor is a ratio of likelihoods and need only be evaluated once for a proposed transition. As the denominator is known from the previous iteration, only the numerator need be evaluated. In the context of registration, this part corresponds to the image term and would involve costly computations if evaluated repeatedly. Note also that for basis functions with compact support (or approximately so), only part of the image term need be updated to evaluate the ratio. The ratio on the right-hand side and the ratio of proposals are simply decomposed over the sequence of previously defined basic moves, then efficiently evaluated using Eq. (6), (13), (14) and expressions similar to Eq. (9), (10). For the latter, statistics $\kappa_k$ are kept up to date (for all bases) using efficient rank one updates derived in [9]. Alternatively, the necessary statistic $\kappa_k$ can be recomputed from scratch only for the bases under consideration. This is usually much more efficient (cf. algorithmic complexity in II-C5).

Each transition kernel $P_k$ satisfies a detailed balance condition. In terms of these transition kernels, the MCMC chain proceeds as follows. Random variables $k_1, k_2, \ldots$ taking values in $\{1, 2, \ldots, M\}$ are chosen according to some scheme and the corresponding transition kernel $P_{k_t}$ is used at time $t$. Conventional schemes include the random-scan, where the $\{k_t\}$ are i.i.d uniform, and the deterministic scan that cycles through $\{1, 2, \ldots, M\}$ in natural order (see *e.g.* [20]). For the random scan, the global transition kernel also satisfies detailed balance conditions. For both schemes, the MCMC chain has stationary distribution $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})$ after incorporating collapsed Gibbs updates of $\boldsymbol{c}$. Highlights of the MCMC scheme main constituents are summarized in Fig. 10.

*4) Markov chain mixing improvement:* Similarly to Gibbs sampling of temperature parameters, Gibbs sampling of voxel GMM assignments $\boldsymbol{c}$ within updates separated from those of $\boldsymbol{w}, \boldsymbol{z}$ potentially hampers the mixing of the Markov chain for any finite, practical duration of the MCMC run. If at any point in time, a data point that should be regarded as an outlier (*e.g.* an image artifact), or a group of such points, is assigned to a 'non-outlier' mixture component, the disjoint sampling generally causes the chain to remain stuck in the vicinity of the corresponding local mode of the posterior $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})$: the desired reverse assignment move virtually occurs with probability zero after readjustment of $\boldsymbol{w}, \boldsymbol{z}$. This defect is critical as such failure scenarii happen with overwhelming probability. Fortunately, joint proposals for $\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}$ can be designed at little cost, even more so after noting that the component-wise proposals for $w_k$ (Eq. (13), (14)) and $z_k$ only *indirectly* depend on $\boldsymbol{c}$. The transition $Q_k(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{c}})$ proceeds in two steps. First, $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$ is proposed as per Algorithm 1. Then, $\tilde{\boldsymbol{c}}$ is sampled by component-wise collapsed Gibbs sampling of each $\tilde{c}_i \sim p(\tilde{c}_i \,|\, \tilde{c}_{j<i}, c_{j>i}, \tilde{\boldsymbol{w}}, \mathcal{D}, \mathcal{H})$ in turn. For efficiency, only the subset of voxels in the support of updated basis functions is sampled, and voxel assignments are updated only once in case of overlapping supports. The two-step move is accepted or rejected based on the acceptance ratio (15), replacing $\boldsymbol{c}$ by $\tilde{\boldsymbol{c}}$ where necessary. The order of voxel

traversal is reversed according to the state of reverse_traversal. Sampling $\tilde{\boldsymbol{c}}$ and computing its contribution to the acceptance ratio exclusively involves the residuals $r_i$ and $\tilde{r}_i$ of updated voxels prior and after the update $\boldsymbol{w}, \boldsymbol{z} \rightarrow \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$, which were already required to compute the likelihood change in Eq. (16).

*5) Algorithmic complexity:* The algorithmic complexity associated to a transition kernel $P_k$ (proposal and acceptance-reject) is $\mathcal{O}(|\mathcal{S}| \cdot |\mathcal{I}_+| + \sum_{l \in \mathcal{I}_+} V_l + L\,C)$, noting $\mathcal{I}_+$ the set of updated bases, $V_l$ the number of voxels in the support of basis $\phi_l$ and $C$ the number of voxels whose assignments $\tilde{c}_i$ are resampled. The first term includes part of the cost of the proposal $\boldsymbol{w}, \boldsymbol{z} \rightarrow \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$ and its impact on the ratio of prior probabilities. The second term is replicated three times and can be heavily parallelized in each case: once to compute $\mu^l_{\text{pos}}, \Sigma_l$ in Eq. (13), (14) for $l \in \mathcal{I}_+$, twice to evaluate and store differences in the displacement fields (resp. residual images) over the support of basis functions in $\mathcal{I}_+$ following their update. The last term accounts for all computations related to resampled voxel GMM assignments $\tilde{\boldsymbol{c}}$. When a move that involves the inclusion or removal of a basis function from the active set is accepted, an additional $\mathcal{O}(|\mathcal{S}|^2 + M \cdot |\mathcal{S}|)$ cost is involved to maintain statistics $\kappa_k$ over all bases in the dictionary, with the right-hand term being parallelizable into $M$ disjoint $\mathcal{O}(|\mathcal{S}|)$ operations. The $\mathcal{O}(M \cdot |\mathcal{S}|)$ cost upon inclusion or deletion of a basis can be replaced by a $\mathcal{O}(|\mathcal{S}|^2)$ cost per proposed move, which is usually more efficient.

*6) Initialization:* The chain is initialized from the output of the deterministic algorithm presented in [9] which progresses greedily in the space of parameters $\{\boldsymbol{z}, \lambda, P\}$ towards a local maximum of their joint posterior. We comment, however, that any registration algorithm could reasonably be used to initialize the chain.

## III. Predictive Uncertainties: Marginal Likelihood Maximization vs. Exact Inference

The 'sparse Bayesian' model presented in Fig. 1 is inspired by the Spike-&-Slab model of Mitchell and Beauchamp [21] and the Relevance Vector Machine (RVM) proposed by Tipping [22] for tasks of regression and classification. In the latter work, the author approaches the problem of inferring an optimal sparse regression function from the standpoint of Automatic Relevance Determination (ARD). Point estimates of the hyperparameters that govern basis selection (and in fact of all hyperparameters) are sought in a first step by maximizing the marginal likelihood or *evidence* as per Eq. (17):

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) \\ &= \arg\max_{\boldsymbol{\theta}} \int_{\boldsymbol{w}} p(\mathcal{D}|\boldsymbol{w}, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{w}|\boldsymbol{\theta}, \mathcal{H}) d\boldsymbol{w} \end{aligned} \tag{17}$$

where $\boldsymbol{\theta} = \{\boldsymbol{z}, P, \lambda\}$ using our notations. If non-uniform, proper hyperpriors on $\boldsymbol{\theta}$ are assumed, $\boldsymbol{\theta}^*$ maximizes the posterior $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) \propto p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta}|\mathcal{H})$ instead. In a second step, the distribution of weights $w_k$ is characterized conditionally to the selected model,

$$p(\boldsymbol{w}|\mathcal{D}, \mathcal{H}) \approx p(\boldsymbol{w}|\boldsymbol{\theta}^*, \mathcal{D}, \mathcal{H})\,. \tag{18}$$

This strategy is typically successful in reaching strongly sparse solutions with good predictive power but, above all else, is motivated by its computational efficiency. Dedicated schemes relying on linear algebra and rank one updates make it possible to efficiently, iteratively build the set $|\mathcal{S}|$ of relevant basis functions $\phi_k$ from scratch. See for instance [23], and [9] for an extension to the wider family of priors required for registration tasks. The approximation of Eq. (18) is justified by observing that the full posterior $p(\boldsymbol{w}|\mathcal{D}, \mathcal{H})$ is obtained by summing over all conditional posteriors $p(\boldsymbol{w}|\boldsymbol{\theta}, \mathcal{D}, \mathcal{H})$, conditioned on the value $\boldsymbol{\theta}$, weighted by the posterior probability $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$ for this value:

$$p(\boldsymbol{w}|\mathcal{D}, \mathcal{H}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{w}|\boldsymbol{\theta}, \mathcal{D}, \mathcal{H}) p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) d\boldsymbol{\theta}\,. \tag{19}$$

Now if the available data $\mathcal{D}$ is informative enough, $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$ will be sharply peaked around its mode(s). In the limit case where $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$ is a Dirac centered at its single mode $\boldsymbol{\theta}^*$, Eq. 18 is retrieved exactly, and the two-step scheme outlined in Eq. (17), (18) is justified. Moreover in the case of sparsity governing parameters $\boldsymbol{z} = \begin{pmatrix} z_1 & \cdots & z_M \end{pmatrix}^\intercal$, Tipping [22] argues that, even if several combinations of parameters are highly probable due to the presence of redundant functions $\phi_k$ in the dictionary of bases, they should roughly lead to the same optimal solution $\mathrm{u}^*$ and an approximate mode (or the expectation) of $p(\mathrm{u}|\mathcal{D}, \mathcal{H})$ should still be correctly evaluated. Regardless, we now demonstrate why this evidence-based approximation will typically fail to properly approximate higher order moments of the full posterior, resulting for instance in poor approximation of the real predictive uncertainty. There are two main breakdown situations for the evidence-based approximation of the full posterior assumed in Eq. 18.

Firstly in absence of data, the assumption that the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$ of hyperparameters is well approximated by a Dirac collapses. Indeed the posterior then resembles the prior distribution $p(\boldsymbol{\theta}|\mathcal{H})$, which is typically flat. This scenario is relevant in the case of basis selection parameters $z_k$, since associated basis functions $\phi_k$ have a local support over which reliable data may be missing. Away from data and without strong incentive to include the basis to increase the deformation regularity, the probability of basis inclusion (resp. exclusion) is $\pi_k$ (resp. $1 - \pi_k$), and for neutral values of $\pi_k$, the choice of excluding the basis is arbitrary.

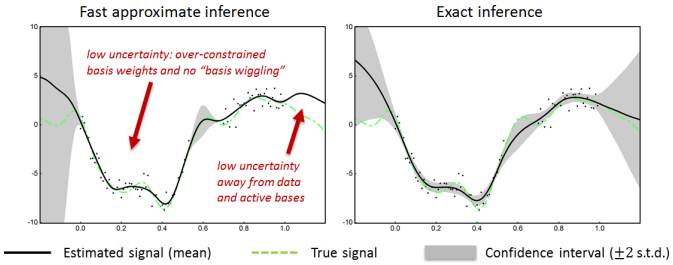


Fig. 4. Comparison of approximate evidence-based inference and faithful MCMC inference for the sparse Bayesian model, on a $1D$ regression task. Data points (black dots) are sampled with additive $i.i.d.$ Gaussian noise from the true signal (dashed green line). The consistence of the fast and faithful estimates of the regressor function (black lines) is satisfactory (w.r.t. uncertainty levels), even more so in the presence of data. Esimates of uncertainty (grey ribbon), however, can be inconsistent.

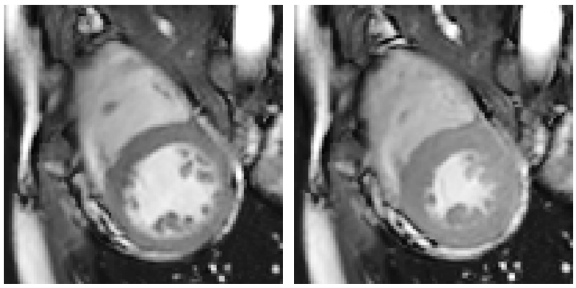

Fig. 5. Registration setting: (left) fixed image, and (right) moving image, at resolution $1.25\text{mm} \times 1.25\text{mm}$.

Secondly and even in the presence of data, many combinations of active bases could have quasi-identical probability. When using radial basis functions for instance, the location of basis centers can be slightly perturbed without significantly affecting the posterior probability of the new configuration. The optimal value of basis weights $\boldsymbol{w}$ under two such perturbations will slightly differ however, as well as the resulting transformation $\Psi$. The evidence-based approximation of Eq. (18) relies on a single – perhaps only marginally superior – configuration, whereas the true posterior sums over all such configurations, as seen from Eq. (19). As it turns out, 'basis wiggling' accounts for a significant part of the uncertainty.

## IV. Experiments and Results

The following experiments aim to qualitatively evaluate the consistency of posterior distributions inferred by the Variational Bayes approximate inference scheme and the MCMC asymptotically exact inference scheme.

### A. Material & Experimental Setting

We focus on the 2D registration example of Fig. 5. For the approximate-based inference, the methodology of [9] is used without change. The multiscale dictionary hence uses Gaussian RBFs at three different scales (isotropic, $\sigma = 6\text{mm}$, 12mm and 24mm), for a total of approximately $7 \cdot 10^4$ basis functions, of which no more than $50 - 100$ are typically active at a time (both with VB and MCMC approaches). We set the differential operator $D$ to the Laplacian of the displacement field. The Gaussian Mixture model of intensity residuals has $L = 5$ components: hyperparameter values $\boldsymbol{\beta}_*$ for the proposal distribution learned during the burn-in phase (sec. II-C2) indicate that 2 or 3 components would suffice. No strong dependence of the results on the number of components was observed. All hyperpriors use small uninformative values $a_0 = b_0 = \gamma_0 = \delta_0 = 10^{-10}$, and $\kappa_0 = 0.5$ for a non-informative (Jeffreys) Dirichlet prior. The MCMC chain was run for roughly $7 \cdot 10^5$ transitions and 500 samples were regularly extracted. Approximately $7 \cdot 10^4$ additional samples were discarded as part of the burn-in phase, during which the parameters of the proposal distribution were fine-tuned (cf. section II-C2). The tuning relies on a set of sufficient statistics, such as the average energy and the average voxelwise square intensity residuals per sample. The averages are computed using a scheme that downweights the early samples: a fixed learning rate is initially applied before reverting to a classical (inverse linear) weighting, drawing inspiration from the SAEM scheme of *e.g.* [7]. The free parameter $s$ controlling the spread of proposals compared to the second-order approximation of the posterior (section II-C2) was set to 1 (spread unchanged). The observed acceptance rate varied between 20–45% under sensible variations of the experimental setting, and between 27–34% during the run of interest with the settings described above. Examples of samples are reported in Fig. 6. As an order of magnitude, the run takes 10 minutes on a standard laptop with a naive implementation.

Finally to gain more insight into the behaviour of VB and MCMC approaches, we also experiment with an MCMC chain that proceeds as described above *except* for the choice of active basis functions which, instead of exploring various configurations, is *fixed* to that of the Variational Bayes approach. We refer to this experiment as Fixed Basis MCMC (FBMCMC). The setting is entirely identical to that of the full MCMC, but the only transitions proposed are component-wise updates as opposed to exchange, addition or removal of basis functions.

### B. Results

*1) Naive alternated sampling vs. joint sampling:* Fig. 7 demonstrates the benefit of a careful design of the Markov chain. The left-most figure displays the estimated mean displacement, under the aforementioned experimental setting, if moves in the space of transformation parameters are done separately from the resampling of voxelwise assignments to components of the noise mixture instead of jointly (right-most figure). In this example, a local discrepancy in the intensity profiles of the fixed and moving images induces a spurious maximum in the joint posterior distribution of transformation parameters and voxel labels (cf. section II-C4). A systematic drift towards this mode was observed in all runs where the sampling was performed in an alternated manner, for the whole duration of the run, whereas systematic recovery was observed under the improved scheme. Similar observations were made in experiments where temperature parameters were treated by Gibbs sampling instead of analytically marginalized over.

*2) VB vs. MCMC – estimated displacement:* Fig. 8 reports the mean displacement reported respectively by the evidence-based inference scheme and by the MCMC inference scheme. As anticipated from the discussion of section III, very good agreement between the evidence-based and MCMC-based estimates of the displacement is observed. Upon close inspection,

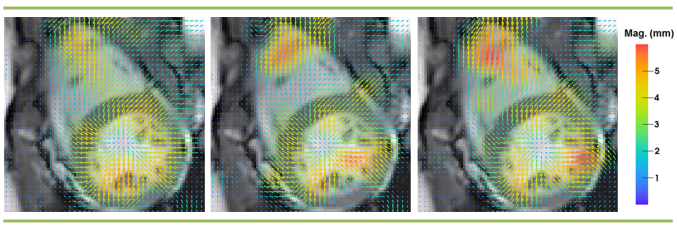


Fig. 6. Three example samples returned by the MCMC run.

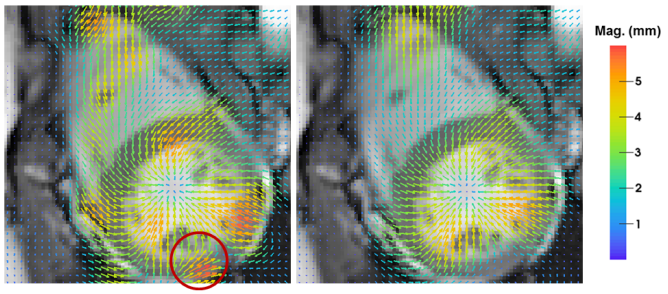


Fig. 7. Comparison of estimates of the posterior mean returned by MCMC characterization. (Left) Sampling activation variables $\boldsymbol{z}$ and corresponding weights $\boldsymbol{w}$, alternatively with voxel mixture assignments $\boldsymbol{c}$. (Right) Joint sampling, as per the approach proposed in section II-C4.

minor differences are noted in some areas with flat intensity profiles or otherwise low confidence (such as that resulting from artefacts, or disagreeing intensities in the fixed and moving image). Their magnitude is lower than the level of uncertainty in the output of registration, as estimated from the MCMC scheme.

*3) VB vs. MCMC – uncertainty estimates:* Fig. 9 compares the estimates of uncertainty obtained from the MCMC characterization of the posterior and those obtained from the Variational Bayes inference. For the MCMC inference, relevant statistics are estimated from the set of samples returned by the run. To study the spatial localization of uncertainty, we visualize at each voxel center $x_i$ the $2\times2$ covariance matrix of the posterior distribution $p(u(x_i)|I,J,\mathcal{H})$ of the corresponding displacement vector $u(x_i)$. This is reasonable under the assumption that the posterior on displacements is approximately mono-modal and Gaussian. The voxelwise covariance matrix, or its square root (homogeneous to a standard deviation), can be visualized as a $2D$ tensor that encodes uncertainty at this point along any direction. Fig. 9 displays the resulting tensor map (bottom row) and a scalar summary (upper row).

On the one hand, the order of magnitude of uncertainties

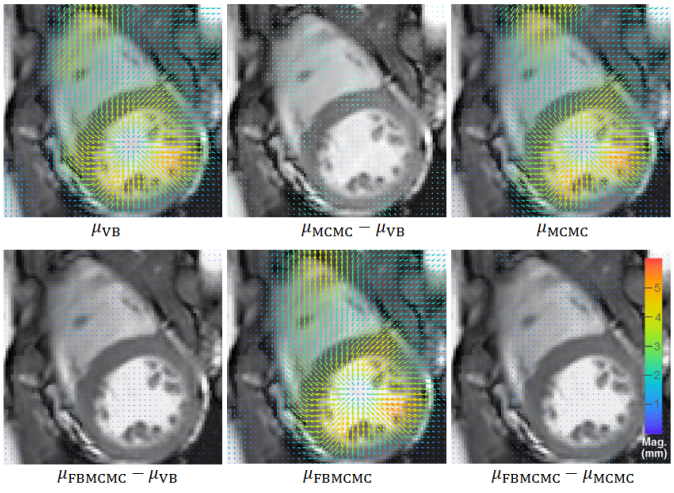


Fig. 8. (Top Row) Comparison of the posterior mean displacement returned by VB (left) vs. MCMC (right), and difference between the two (middle). (Bottom Row) Mean displacement returned by Fixed Basis MCMC (middle) and the difference with the VB (resp. MCMC) estimate (left, resp. right).

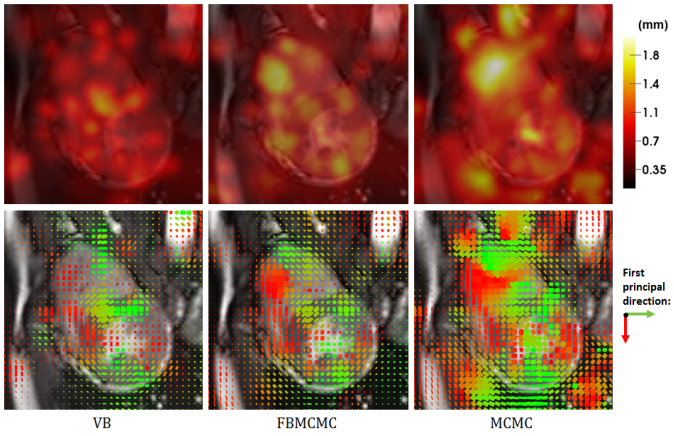


Fig. 9. Estimates of uncertainty obtained by characterizing the posterior distribution of the sparse Bayesian model by (Left column) Variational Bayes (Right column) MCMC sampling (Middle column) Fixed Basis MCMC sampling. (Second row) Tensor visualization of the displacement uncertainty: each tensor encodes the square root of the $2 \times 2$ covariance displacement matrix at this location. Tensor elongation along a direction indicates higher uncertainty along that direction. The color scheme encodes the direction of the first eigenvector. (First row) Trace of the square root covariance.

under the true posterior (typically $\sim 1mm$ for a $95\%$ confidence interval), as estimated by MCMC sampling, is consistent with both the magnitude of the underlying motion (no more than $5mm$, see fig. 8) and the resolution (voxel dimensions: $1.25mm \times 1.25mm$). As expected, uncertainty is higher in regions with little structured content (no intensity gradients) and in the direction of contours. On the other hand, the VB scheme does not appear to reliably approximate the true uncertainty. Its order of magnitude is generally underestimated. Moreover, VB-based uncertainty may lack spatial coherence in regions that are textureless, with a flat intensity profile (*e.g.* in the right ventricle on Fig. 5). This hints at the fact that, when relying on the evidence-based (VB) scheme, regions of high uncertainty are localized nearby the inferred (unique) set of active basis functions.

*4) Fixed Basis MCMC:* Fig. 8 (second row) and Fig. 9 (middle column) report the estimates of the mean and uncertainty for the Fixed Basis MCMC scheme. The estimated mean displacement is in good agreement with both approaches. Moreover the magnitude of the difference between FBMCMC and VB (resp. FBMCMC and MCMC) is generally below that of the residual displacement between VB and MCMC. The FBMCMC approach, similarly to the VB approach, underestimates uncertainty in regions of flat intensity (*e.g.* bottom right of the image) and displays small localized uncertainty peaks. The magnitude of the predicted uncertainty is globally consistent with that of the VB scheme (Fig. 9, similar tensor sizes in the first and second rows), albeit sometimes slightly superior, typically nearby active basis functions (*e.g.* in the anterior part of the right ventricle).

## V. Discussion

### A. Markov Chain design for efficient and reliable inference

The proposed model of registration copes with various unknowns in the image and transformation model: the noise

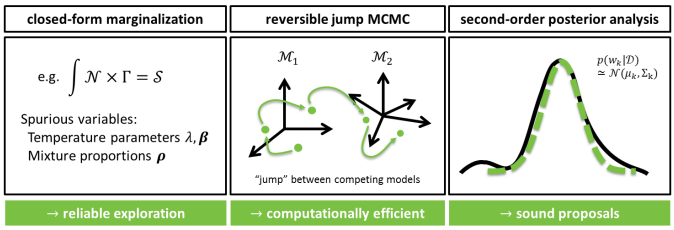


Fig. 10. Main constituents of the MCMC scheme.

level and its spatial variability, the regularity of the hidden motion, and the optimal parametrization of the displacement field itself. This renders inference challenging and special care has to be taken in the design of the Markov chain. Typically several joint configurations of the regularization hyperparameters, the noise levels (and voxelwise mixture assignments) and the displacement field constitute local maxima in the joint parameter space. To prevent the chain from remaining stuck around poor local maxima, it proved useful to analytically marginalize over nuisance variables (noise and regularization levels, mixture proportions) as well as to jointly sample transformation parameters and voxel assignments to mixture components (as opposed to alternate between sampling one or the other).

The reversible jump MCMC framework that we implement has strong connection with the jump diffusion process of Grenander and Miller [24] and the birth-and-death kernel framework. It allows to move freely in the space of transformation parameters but also and conccurently, in the space of admissible parametrizations. By circumventing the costly computation of Bayes factors (ratios of evidence for competing parametrizations), it effectively renders MCMC inference tractable for the sparse Bayesian model of registration, even with large dictionaries of basis functions ($\sim 10^5$ in our experiments).

Full-dimensional moves over the space of transformation parameters were not implemented, as calibrating such transitions calls for the particularly expensive computation of large (non-diagonal) Hessian matrices. This renders them inefficient unless e.g., exploiting dedicated procedures inspired from limited memory quasi-Newton methods [18]. Component-wise transitions are also particularly suitable provided that the set of active bases must be jointly explored.

### B. Variational Bayes and MCMC inference

Experimental results point towards a good empirical correspondence between the mean estimates of displacement returned by the approximate VB inference and (asymptotically) exact MCMC inference, particularly in the presence of informative data. Unfortunately they also evidence limitations of the approximate VB scheme for purposes of uncertainty quantification. This defect is offset by a significantly faster running time for the VB scheme (one order of magnitude).

As shown in section III, VB inferences selects a single parametrization by means of marginal likelihood maximization, although this optimal parametrization will often be only marginally so. Discarding all marginally sub-optimal parametrizations results in generally unreliable estimates of uncertainty. This is also evidenced by the lower magnitude of uncertainty predicted by the fixed basis MCMC scheme. Approximate schemes that circumvent this issue can likely be devised, for instance by keeping track of several sets of relevant explanatory variables [25]. The uncertainty on optimal basis locations could be accounted for in the VB scheme, either in an ad-hoc manner by local perturbations of the basis centers when sampling a transformation from the variational posterior, or in a more formal way by regarding basis centers as random variables whose associated variational posteriors are to be estimated.

The RVM$^*$ basis augmentation strategy of Rasmussen and Candela [26] partially addresses the second issue of uncertainty underestimation in absence of data. It is applicable only when voxelwise estimates of uncertainty are expected, as opposed to full transformation samples. Another strategy would be to relax the form of the variational posterior family so that it can better approximate the true posterior away from data, with the constraint that the computational burden remain suitably low under such a factorization. Alternatively we note instead the high potential for parallelization of the proposed MCMC approach, which could make it more amenable to routine use on real data.

Finally VB makes parametric assumptions about the form of the posterior distribution, and infers hyperparameter distributions whereas the proposed MCMC scheme generally marginalizes over such hyperparameters. This is likely to account for some of the minor differences observed between the VB and Fixed Basis MCMC approaches (sec. IV-B4).

### C. Underlying assumptions of the Sparse Bayesian registration model

The validity of model assumptions may affect the quality of uncertainty estimates. Possible biases stem on the one hand from the inexactness of the generative model of images (modelling the intensity residual as a mixture of Gaussians, discarding spatial correlations between residuals), on the other hand from inexactness of the transformation model (the parametrization of the transformation as well as the choice of prior). Their impact was not thoroughly explored so far, but this work provides the methodological framework to do so.

The assumption that source and target image intensities coincide up to spatially varying noise mostly holds in the context of mono-modal registration. For multi-modal registration, a mapping function between source and target image intensities should be used (as in *e.g.* [27], [28]) and can be regressed within a probabilistic framework [29].

## VI. Conclusion

In this article we explored the properties of the proposed sparse Bayesian model of registration for the purpose of uncertainty quantification. We emphasize the distinction between the Bayesian model itself and inference schemes used to estimate posterior distributions under this model. In previous work [9] an efficient but approximate inference scheme was developed, based on Variational Bayesian arguments and the principle

of marginal likelihood maximization. In the present work we design a reversible jump Markov chain that characterizes the exact posterior arbitrarily well (provided that enough samples can be drawn) and answer the two following questions. Firstly, does the fast approximate scheme provide faithful estimates of expectation and uncertainty? Secondly, is the sparse Bayesian model of registration useful for the purpose of uncertainty quantification? We evidence limitations of the approximate inference scheme for uncertainty quantification, but show that the true posterior distribution itself is meaningful: orders of magnitude for the true uncertainty (as characterized by MCMC sampling) are quantitatively reasonable, the uncertainty is higher in textureless regions and lower in the direction of strong intensity gradients.

## ACKNOWLEDGMENT

Part of this work was funded by the European Research Council through the ERC Advanced Grant MedYMA (2011-291080) on Biophysical Modeling and Analysis of Dynamic Medical Images. The first author was funded by the Microsoft Research – Inria Joint Centre.

## REFERENCES

[1] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[2] J. C. Gee and R. K. Bajcsy, "Elastic matching: Continuum mechanical and probabilistic analysis," *Brain warping*, vol. 2, 1998.

[3] C. Broit, "Optimal registration of deformed images," 1981.

[4] P. Risholm, F. Janoos, I. Norton, A. J. Golby, and W. M. Wells III, "Bayesian characterization of uncertainty in intra-subject non-rigid registration," *Medical image analysis*, vol. 17, no. 5, pp. 538–555, 2013.

[5] I. J. Simpson, J. A. Schnabel, A. R. Groves, J. L. Andersson, and M. W. Woolrich, "Probabilistic inference of regularisation in non-rigid registration," *NeuroImage*, vol. 59, no. 3, pp. 2438–2451, 2012.

[6] I. Simpson, M. Cardoso, M. Modat, D. Cash, M. Woolrich, J. Andersson, J. Schnabel, S. Ourselin, A. D. N. Initiative *et al.*, "Probabilistic non-linear registration with spatially adaptive regularisation," *Medical image analysis*, 2015.

[7] F. J. Richard, A. M. Samson, and C. A. Cuénod, "A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences," *Stat Comput*, vol. 19, no. 4, 2009.

[8] M. Zhang, N. Singh, and P. T. Fletcher, "Bayesian estimation of regularization and atlas building in diffeomorphic image registration," in *Information Processing in Medical Imaging*. Springer, 2013, pp. 37–48.

[9] L. Le Folgoc, H. Delingette, A. Criminisi, and N. Ayache, "Sparse bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrization of displacements," *under revision in Medical Image Analysis*, 2016. [Online]. Available: https://hal.inria.fr/hal-01149544

[10] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[11] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[12] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

[13] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.

[14] S. Chib, "Marginal likelihood from the gibbs output," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1313–1321, 1995.

[15] A. Gelman and X.-L. Meng, "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statistical science*, pp. 163–185, 1998.

[16] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, pp. 341–363, 1996.

[17] M. Girolami and B. Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

[18] Y. Zhang and C. A. Sutton, "Quasi-newton methods for markov chain monte carlo," in *Advances in Neural Information Processing Systems*, 2011, pp. 2393–2401.

[19] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[20] G. O. Roberts and J. S. Rosenthal, "Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains," *The Annals of Applied Probability*, pp. 2123–2139, 2006.

[21] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[22] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[23] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse bayesian models," in *Workshop on artificial intelligence and statistics*, vol. 1, no. 3. Jan, 2003.

[24] U. Grenander and M. I. Miller, "Representations of knowledge in complex systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 549–603, 1994.

[25] P. Schniter, L. C. Potter, and J. Ziniel, "Fast bayesian matching pursuit," in *Information Theory and Applications Workshop, 2008*. IEEE, 2008, pp. 326–333.

[26] C. E. Rasmussen and J. Quinonero-Candela, "Healing the relevance vector machine through augmentation," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 689–696.

[27] M. E. Leventon and W. E. L. Grimson, "Multi-modal volume registration using joint intensity distributions," in *Medical Image Computing and Computer-Assisted InterventationMICCAI98*. Springer, 1998, pp. 1057–1066.

[28] A. Guimond, A. Roche, N. Ayache, and J. Meunier, "Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 1, pp. 58–69, 2001.

[29] F. Janoos, P. Risholm, and W. Wells III, "Bayesian characterization of uncertainty in multi-modal image registration," *Biomedical Image Registration*, pp. 50–59, 2012.